Градиентные методы обучения

Повторение

Функция потерь для регрессии

- Еще один вариант – средняя абсолютная ошибка (mean absolute error, MAE)

$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}|a(x_i) - y_i|$$

- Слабее штрафует за выбросы

- Зануляет признаки перед незначищими признаками

Линейная регрессия в векторном виде

$$a(x) = < w, x >$$

- Среднеквадратичная ошибка и задача обучения:

$$\frac{1}{l}\sum_{i=1}^{l}(< w, x_i > - y_i)^2 \to \min\_w$$

Модель линейной регрессии

Среднеквадратичная ошибка и задача обучения:

$$\frac{1}{l}\sum_{i=1}^{l}(< w, x_i > \quad -y_i)^2 \rightarrow \min\_w$$

Отклонения прогнозов от ответов:

$$Xw - y = \begin{pmatrix} \langle w, x_1 \rangle - y_1 \\ \vdots \\ \langle w, x_\ell \rangle - y_\ell \end{pmatrix}$$

Вычисление ошибки

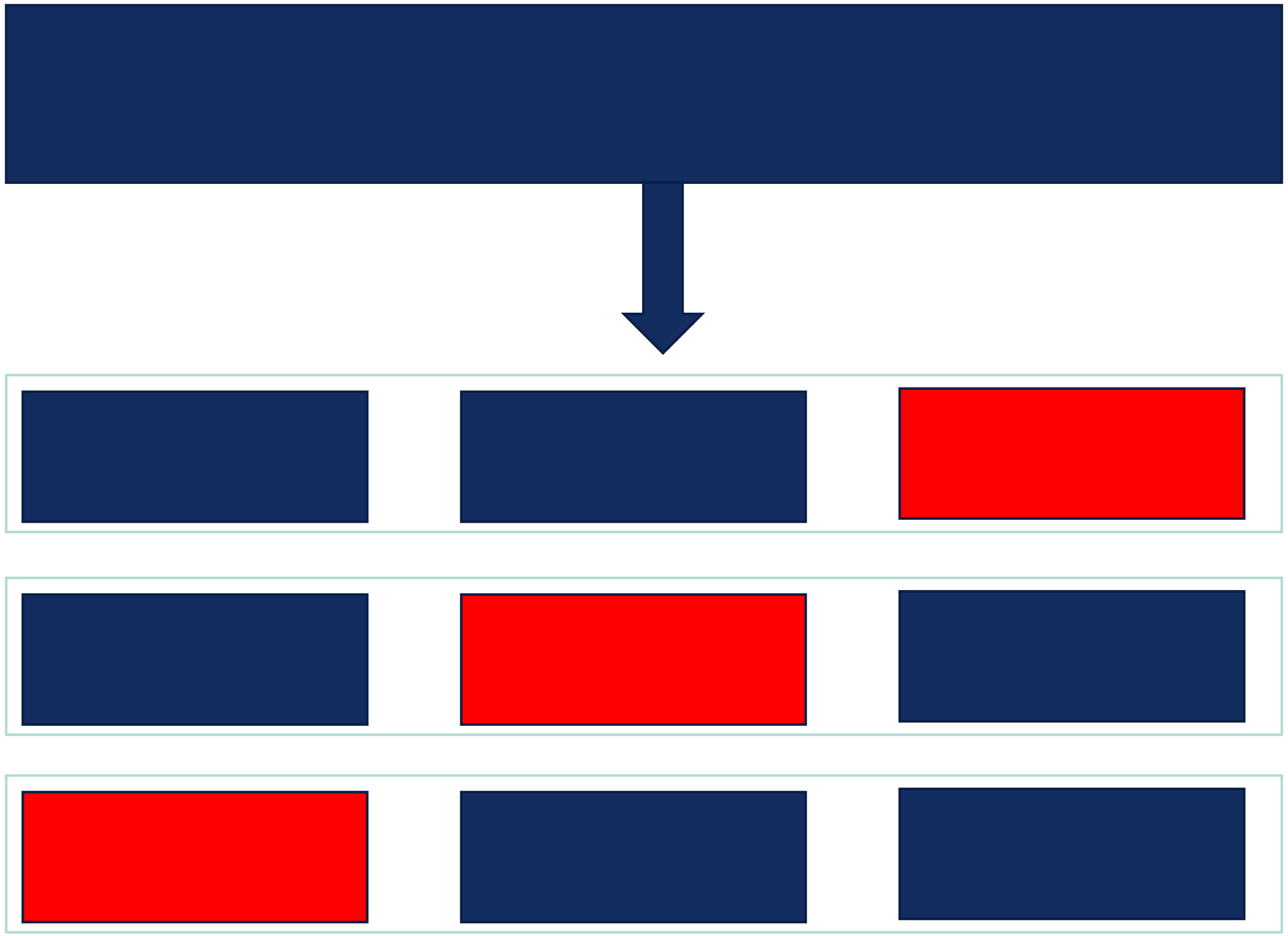
Отклонения прогнозов от ответов:

$$||z|| = \sqrt{\sum_{j=1}^{n} z_j^2}$$

Среднеквадратичная ошибка:

$$\frac{1}{l}||Xw - y||^2 = \frac{1}{l}\sum_{i=1}^{l}(< w, x_i > \; -y_i)^2$$

Кросс-валидация



Признаки переобученной модели

- Большая разница в качестве на тренировочных и
  и тестовых данных(модель подгоняется под тренировочные
  данные и не может найти истинная зависимость)

- Большие значения параметров (весов) w_j модели

Регуляризаторы

- $||z||_2 = \sqrt{\sum_{j=1}^{d} z_j^2} - \mathrm{L}_2 - \text{норма}$
- $||z||_1 = \sum_{j=1}^{d} |z_j| - \mathrm{L}_1 - \text{норма}$

Гиперпараметры

- Гиперпараметр алгоритма — то, что задается вручную
- Пример: коэффициент регуляризации
- Параметр алгоритма — то, что определяется моделью
- Пример: веса регрессии
- Как подбирать гиперпараметры алгоритма?
- Нельзя подбирать по обучающей выборке — это приведет к переобучению
- Нужно использовать дополнительные данные (валидация)

Чуть больше терминов

- После подбора всех гиперпараметров стоит проверить на совсем новых данных, что модель работает

- Обучающая выборка — построение модели

- Валидационная выборка — подбор гиперпараметров модели

- Тестовая выборка — финальная оценка качества модели

Масштабирование признаков

- Отмасштабируем j-й признак
- Вычисляем среднее и стандартное отклонение признака на обучающей выборке:

$$\mu_j = \frac{1}{l}\sum_{i=1}^{l} x_i^j$$

$$\sigma_j = \sqrt{\frac{1}{l}\sum_{i=1}^{l} \left(x_i^j - \mu_j\right)^2}$$

Среднеквадратичная ошибка

- MSE для линейной регрессии:
  $\frac{1}{l}\left||Xw\ - y|\right|^2 \rightarrow \min\_w$

$Q(w_1, \ldots, w_d) = \frac{1}{l}\sum_{i=1}^{l}(w_1 x_1 + \ \ldots + w_d x_d\ - y_i)^2$

Можно посчитать градиент MSE:

- $\nabla\frac{1}{l}\left||Xw\ - y|\right|^2 = \frac{2}{l}X^T(Xw\ - y)$

Приравниваем нулю и решаем систему линейных уравнений:

- $w = (X^T X)^{-1} X^T y$

Аналитическое решение

$$w = (X^T X)^{-1} X^T y$$

- Если матрица $(X^T X)$ вырожденная, то будут проблемы
- Даже если она почти вырожденная, все равно будут проблемы
- Если признаков много, то придется долго ждать

- Обращение матрицы – сложная операция ($O(N^3)$) от числа признаков

- Если заменить среднеквадратичный функционал на другой, то скорее всего не найдем аналитическое решение

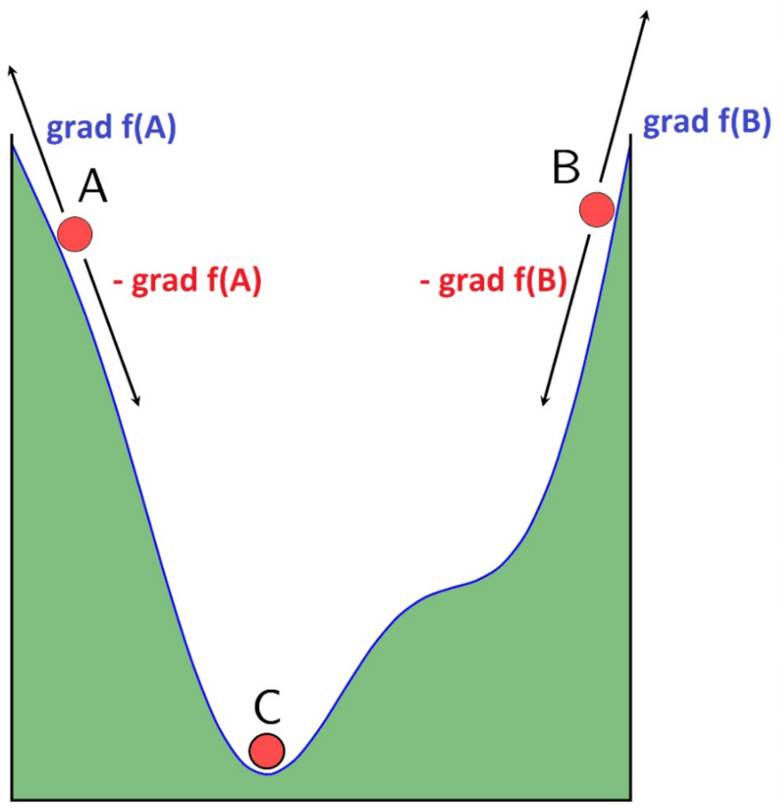
Регуляризация

- Регуляризованный функционал
$$\frac{1}{l}\sum_{i=1}^{l}(< w, x_i > \ - y_i)^2 + \lambda||w||^2 \to \min\_w$$
- Аналитическое решение:
$$w = (X^T X + \lambda I)^{-1} X^T y$$
- Гребневая регрессия (Ridge regression)

Теорема о градиенте

Теорема. Градиент – это вектор, в направлении которого функция быстрее всего растет.

Антиградиент (вектор, противоположный градиенту) – Вектор, в навправлении которого функция быстрее всего убывает.



Метод градиентного спуска

- Наша задача при обучении модели – найти такие веса w, на которых достигается минимум функции ошибки.

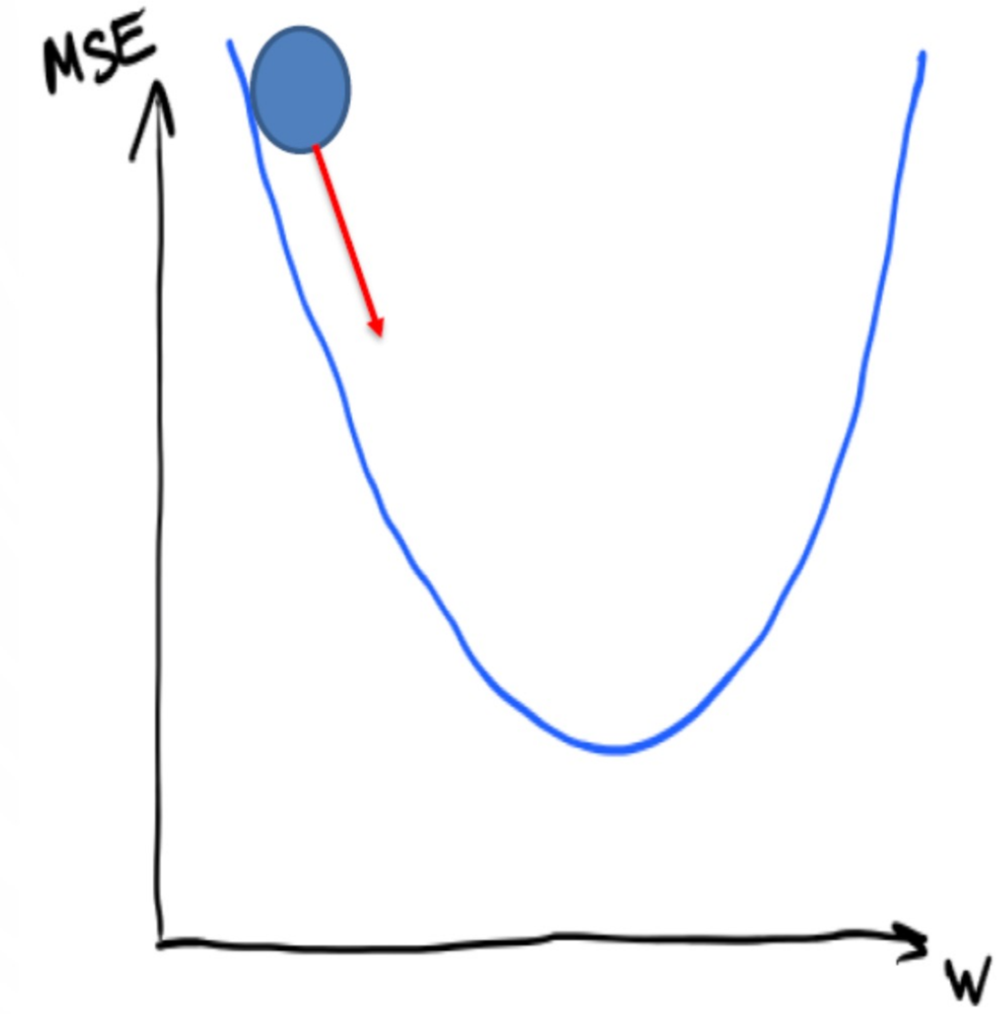
- Грубо говоря, график MSE – парабола

- Идея метода градиентного спуска.

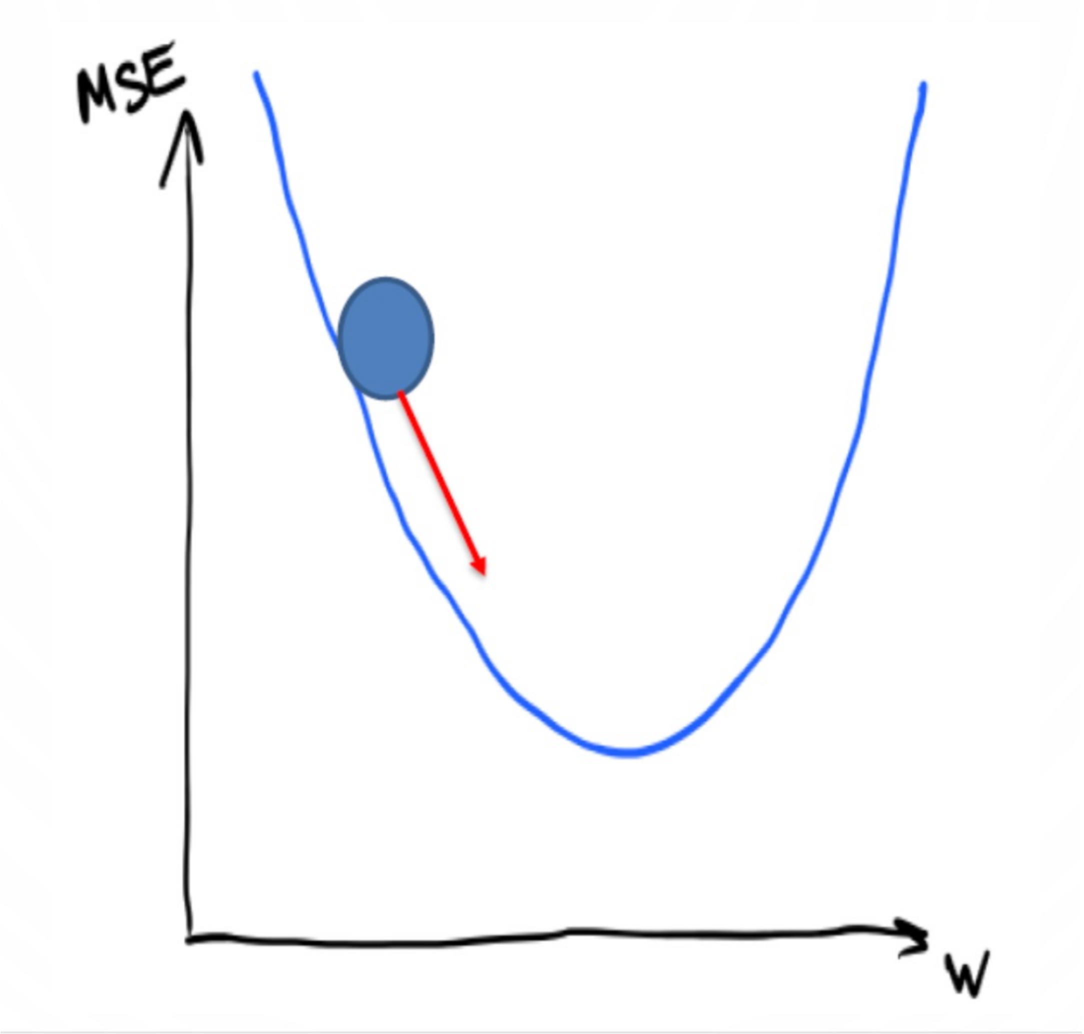
- На каждом шаге движемся в сторону антиградиента функции потерь!

- Вектор градиента функции потерь обозначают $grad\ Q$ или ∇Q

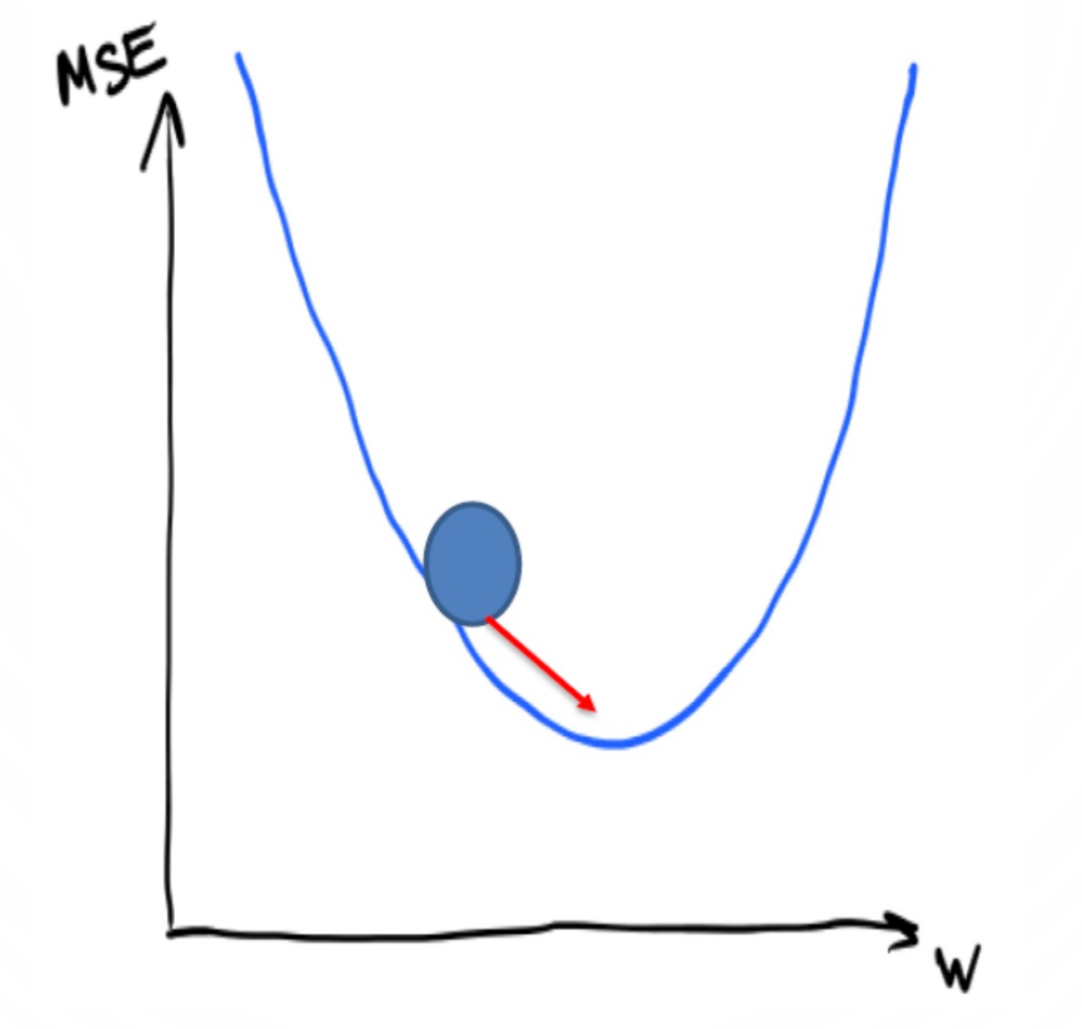
Метод градиентного спуска



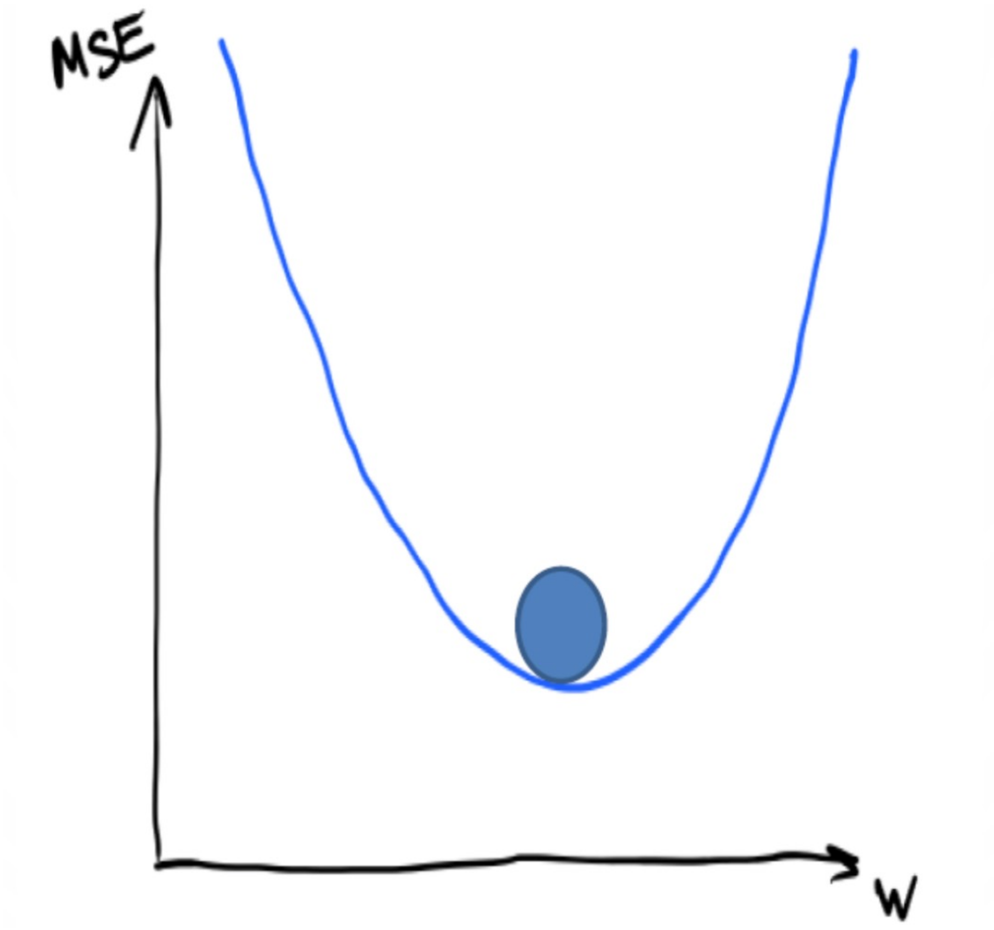
Метод градиентного спуска



Метод градиентного спуска

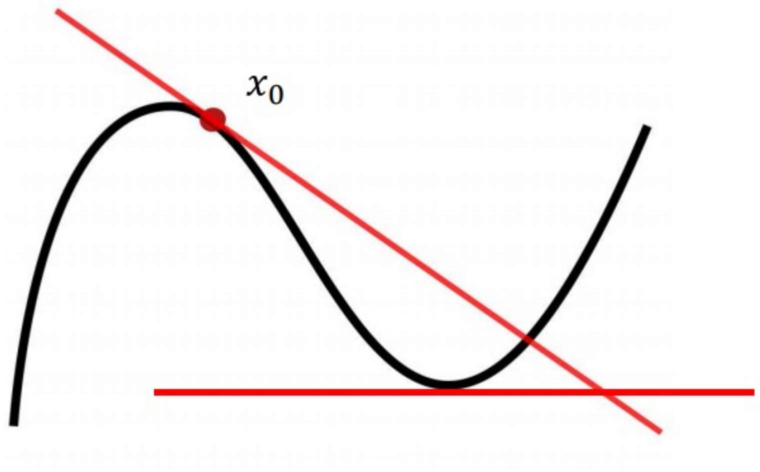


Метод градиентного спуска



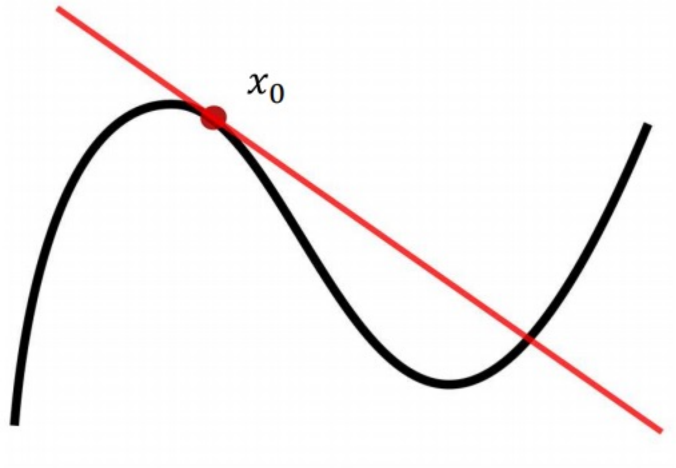
Производная

Если точка x_0 — экстремум и в ней существует производная, то $f'(x_0) = 0$



Производная

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$$

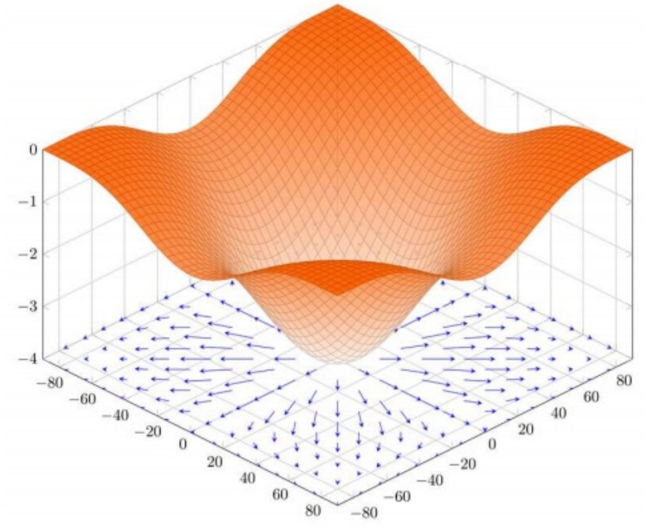


Градиент

- Градиент – вектор частных производных

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d}\right)$$

- У градиента есть важное свойство!
- Зафиксируем точку x_0
- В какую сторону функция быстрее всего растет?



Важное свойство

- Зафиксируем точку x_0
- В какую сторону функция быстрее всего растет?
- В направлении градиента!
- А быстрее всего убывает в сторону антиградиента

Условие экстремума

- Если точка x_0 - экстремум и в ней существует производная, то $\nabla f(x_0) = 0$

- Если функция выпуклая, то экстремум один

- MSE для линейной регрессии — выпуклая!

Градиентный спуск

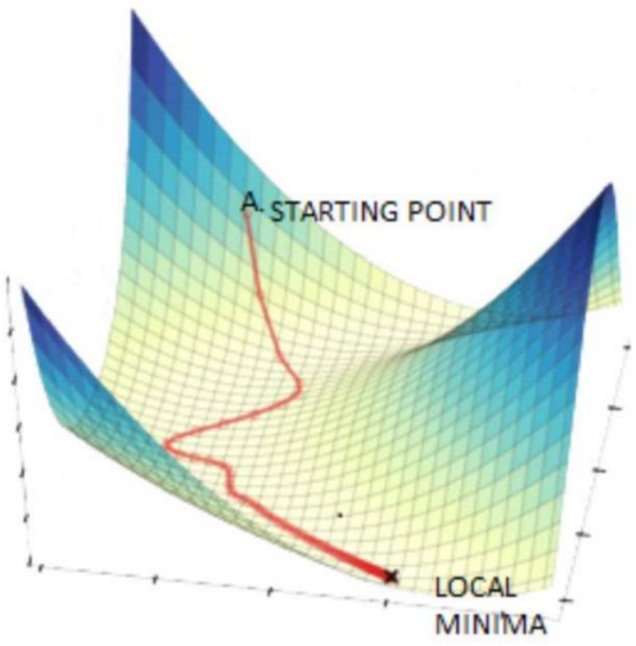
- Если точка x_0 - экстремум и в ней существует производная, то $\nabla f(x_0) = 0$

- Если функция выпуклая, то экстремум один

- MSE для линейной регрессии — выпуклая!

Как это пригодится?

Как это пригодится?



Градиентный спуск

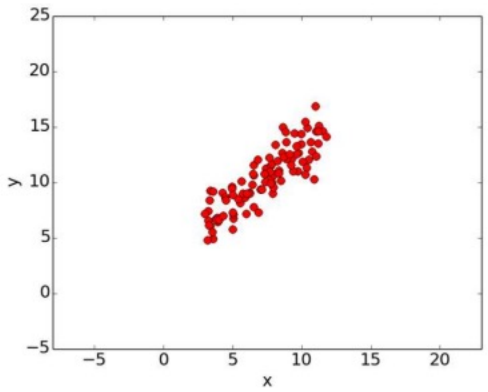
- Стартуем из случайной точки
- Сдвигаемся по антиградиенту
- Повторяем, пока не окажемся в точке минимума

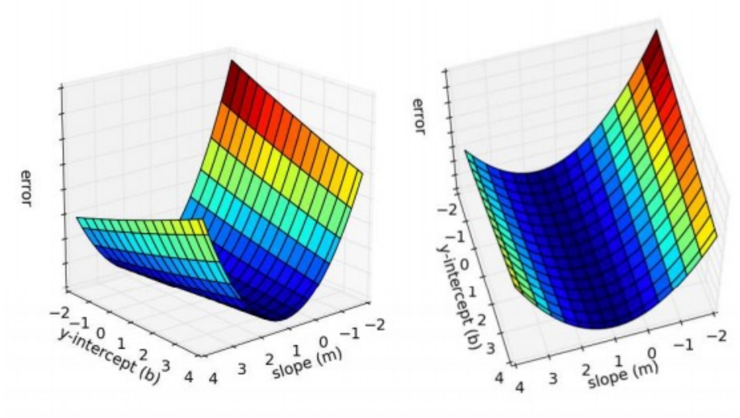
Парная регрессия

- Простейший случай: один признак
- Модель: $a(x) = w_1 x + w_0$
- Два параметра w_1, w_0
- w_1 - тангенс угла наклона
- w_0 - где прямая пересекает ось ординат
- Функционал:

$$Q(w_0, w_1) = \frac{1}{l}\sum_{i=1}^{l}(w_1 x_i + w_0 \ - y_i)^2$$

Парная регрессия





Выборка

Функционал ошибки

https://spin.atomicobject.com/2014/06/24/gradient-descent-linear-regression/

Парная регрессия

$$Q(w_0, w_1) = \frac{1}{l}\sum_{i=1}^{l}(w_1 x_i + w_0 \ - y_i)^2$$

$$\frac{\partial Q}{\partial w_1} = \frac{2}{l}\sum\nolimits_{i=1}^{l} x_i(w_1 x_i + w_0 \ - y_i)$$

$$\frac{\partial Q}{\partial w_0} = \frac{2}{l}\sum\nolimits_{i=1}^{l} (w_1 x_i + w_0 \ - y_i)$$

$$\nabla Q(w) = (\frac{2}{l}\sum\nolimits_{i=1}^{l} x_i(w_1 x_i + w_0 - y_i), \frac{2}{l}\sum\nolimits_{i=1}^{l}(w_1 x_i + w_0 \ - y_i))$$

Линейная регрессия

$$Q(w) = \frac{1}{l}\sum_{i=1}^{l}(< w, x > \ -y_i)^2$$

$$\frac{\partial Q}{\partial w_1} = \frac{2}{l}\sum_{i=1}^{l} x_{i1}(< w, x > \ -y_i)$$

$$\ldots$$

$$\frac{\partial Q}{\partial w_d} = \frac{2}{l}\sum_{i=1}^{l} x_{id}(< w, x > \ -y_i)$$

$$\nabla Q(w) = \frac{2}{l}X^T(Xw \ - y)$$

Начальное приближение

w_0 - инициализация весов

Например, из стандартного нормального распределения

Градиентный спуск

Повторять до сходимости:

$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

Градиент в предыдущей точке

Новая точка

Длина шага

Длина шага

Повторять до сходимости:

$$w^t = w^{t-1} - \boldsymbol{\eta} \nabla Q(w^{t-1})$$

Позволяет контролировать скорость обучения

Метод градиентного спуска

На каждом шаге (на каждой итерации метода) движемся в сторону антиградиента Функции потерь!

- Инициализируем веса $w_0^{(0)}, w_1^{(1)}, w_2^{(0)}, \dots, w_n^{(0)}$

- На каждом следующем шаге обновляем веса, сдвигаясь в направлении антиградиента функции потерь Q:

$$w_0^{(k)} = w_0^{(k-1)} - \nabla Q\left(w_0^{(k-1)}\right),$$
$$w_1^{(k)} = w_1^{(k-1)} - \nabla Q\left(w_1^{(k-1)}\right),$$
$$\dots$$
$$w_n^{(k)} = w_n^{(k-1)} - \nabla Q\left(w_n^{(k-1)}\right),$$

Сходимость

- Начальное приближение: w^0

- Повторять
  $w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$

- Останавливаем если
  $||w^t - w^{t-1}|| < \varepsilon$

Градиентный спуск

Останавливаем процесс, если

$$||w^t - w^{t-1}|| < \varepsilon$$

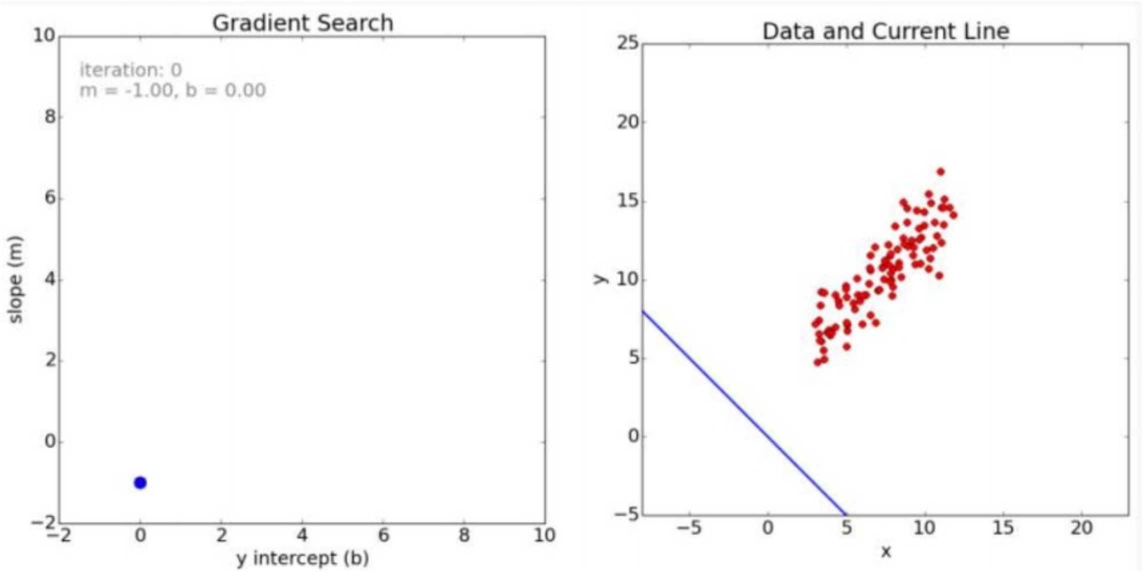
Другой вариант;

$$||Q(w^t) - Q(w^{t-1})|| < \varepsilon$$

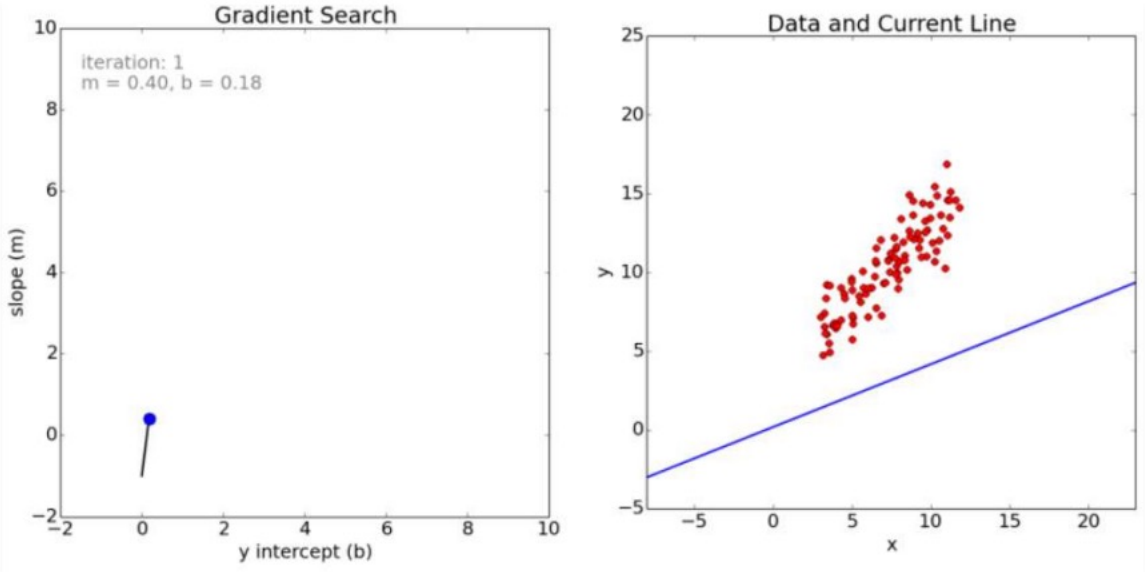
Другой вариант:

$||\nabla Q(w^t)||$ < ε

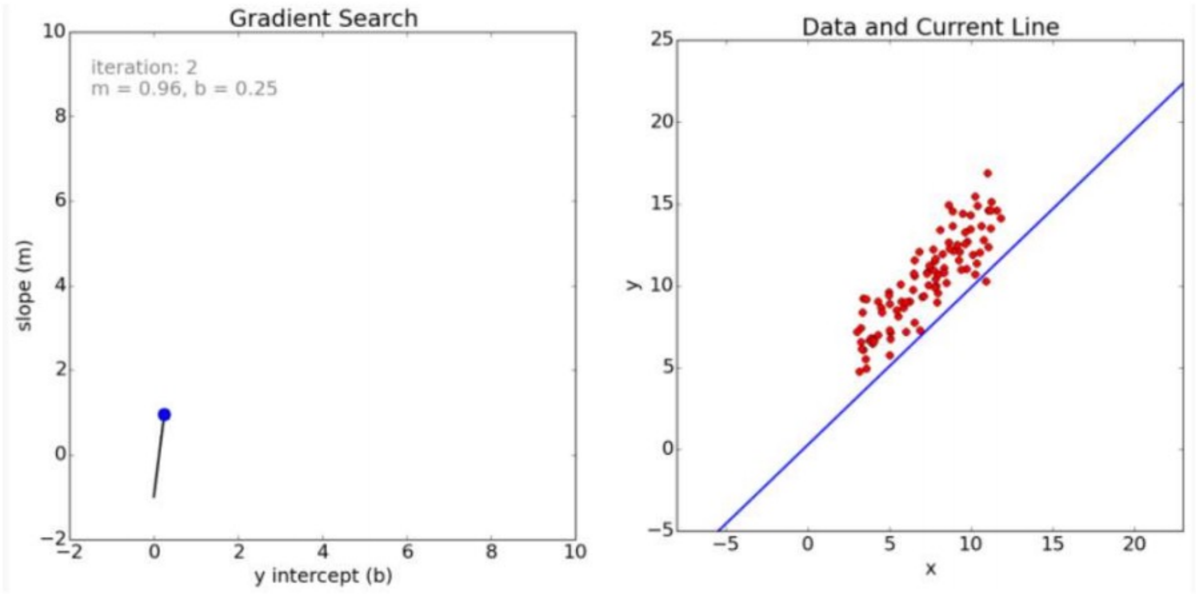
Парная регрессия



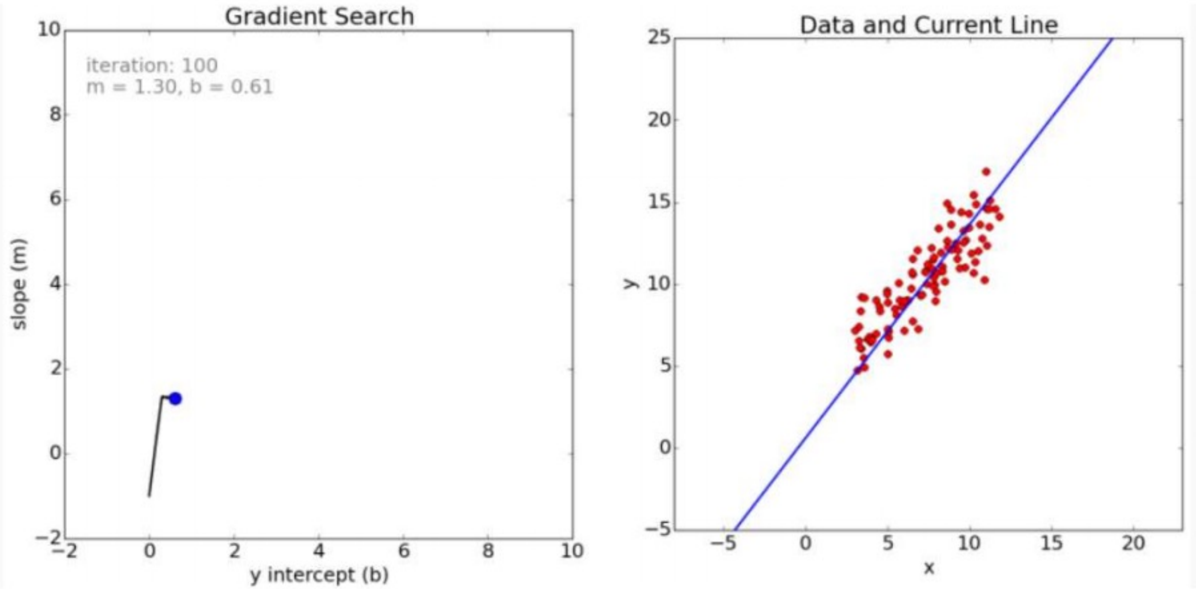
Парная регрессия



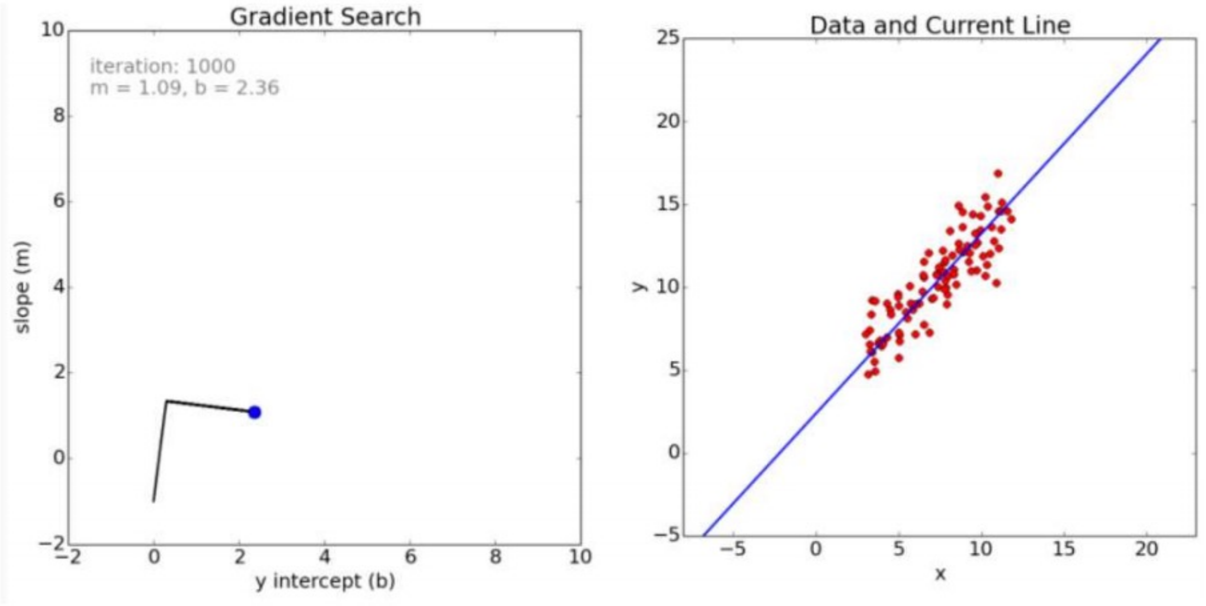
Парная регрессия



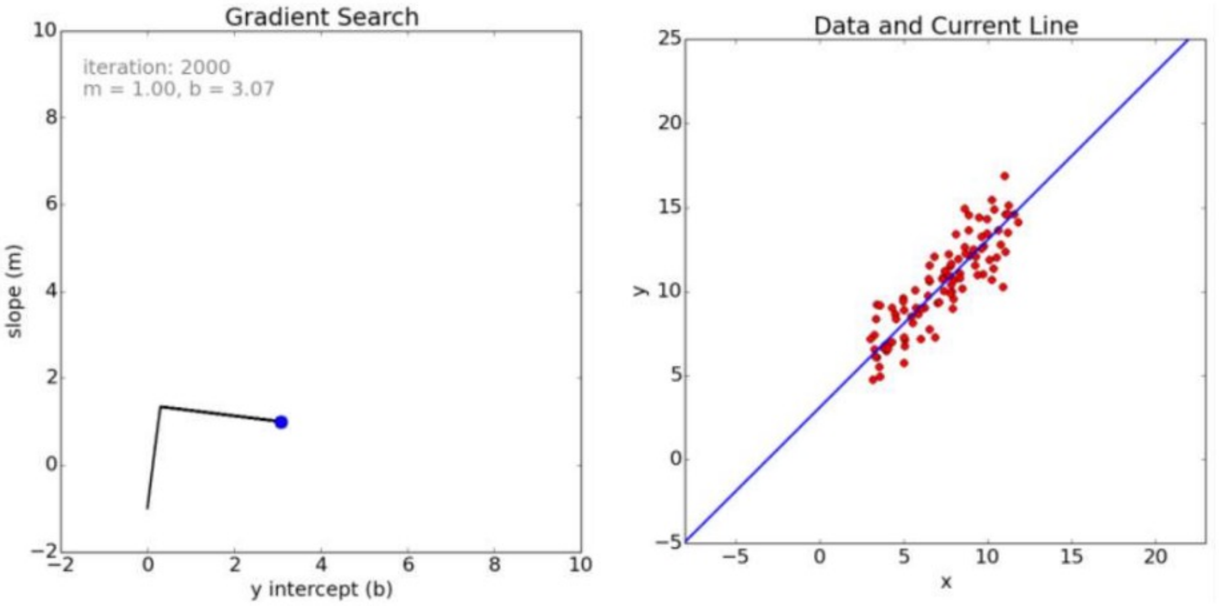
Парная регрессия



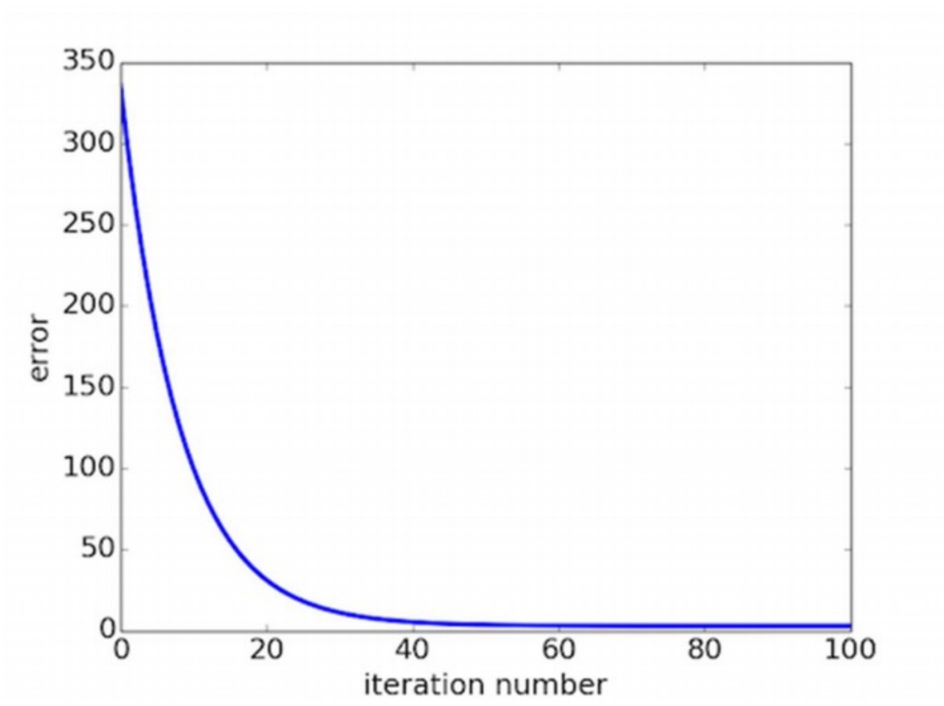
Парная регрессия



Парная регрессия

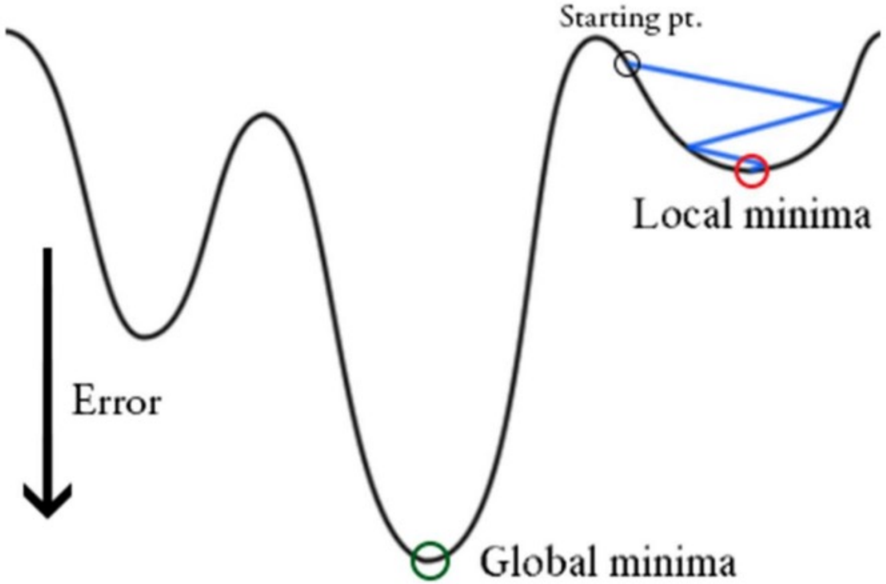


Функционал ошибки

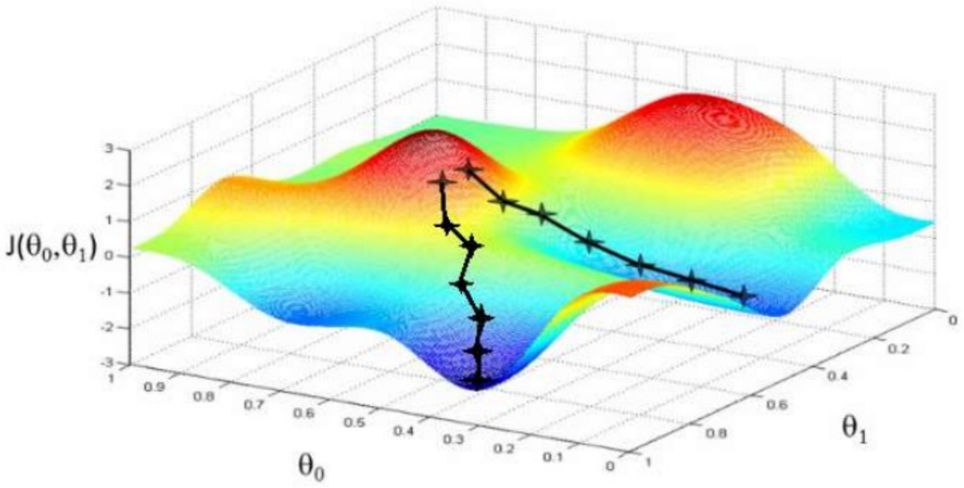


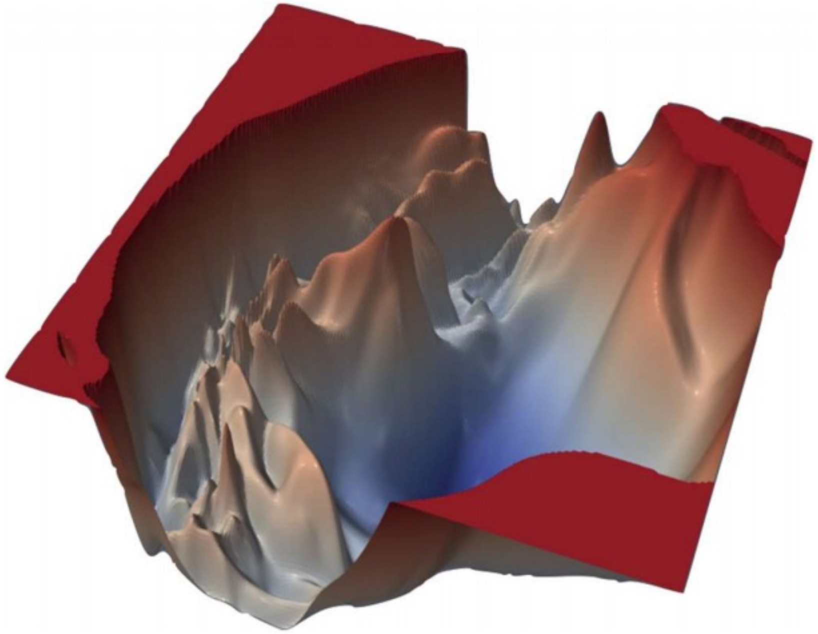
Локальные минимумы

Градиентный спуск находит только локальные минимумы



Локальные минимумы



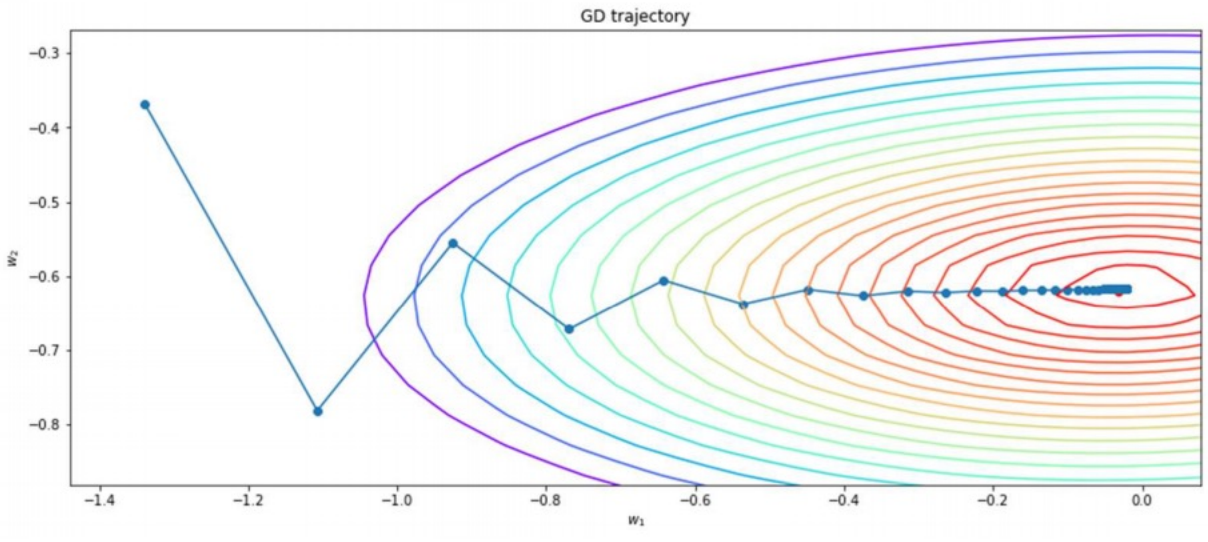


Длина шага

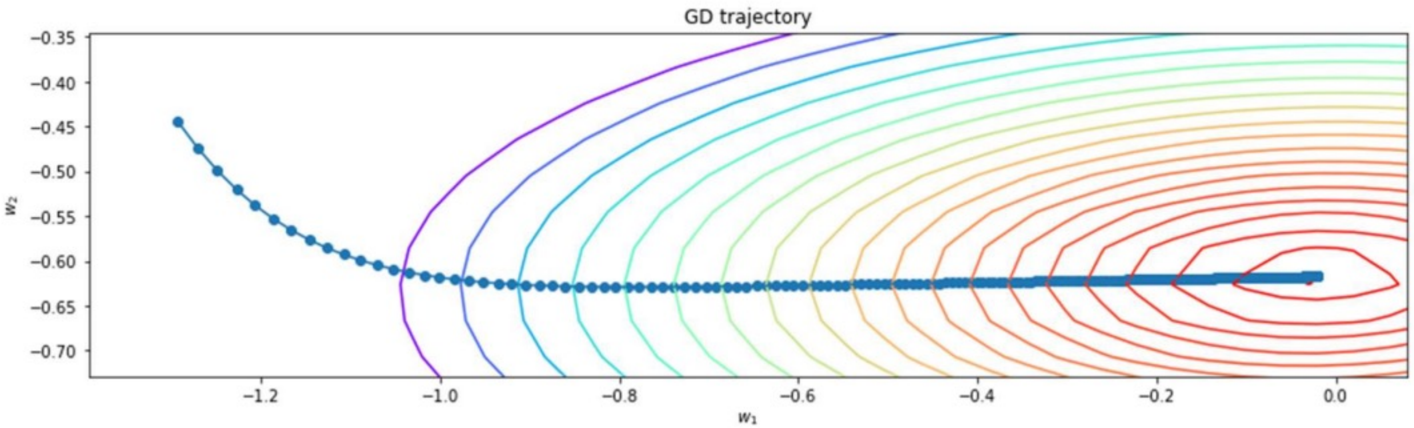
$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

- Позволяет контролировать скорость обучения
- Если сделать длину шага недостаточно маленькой, градиентый спуск может разойтись
- Длина шага – гиперпараметр, который нужно подбирать

Длина шага



Длина шага



GD trajectory
−0.35
−0.40
−0.45
−0.50
−0.55
−0.60
−0.65
−0.70
w_2
−1.2
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
w_1

Переменная длина шага

$$w^t = w^{t-1} - \eta_t \nabla Q(w^{t-1})$$

- Длину шага можно менять в зависимости от шага
- Например: $\eta_t = \frac{1}{t}$
- Шаг наискорейшего спуска:
  $\eta_t = argmin_n Q(w^t) = argmin_\eta Q(w^{t-1} - \eta \nabla Q(w^{t-1}))$

Градиентный спуск

1. Начальное приближение : w^0

2. Повторять:

$$w^t = w^{t-1} - \eta_t \nabla Q(w^{t-1})$$

3. Останавливаемся, если

$$||w^t - w^{t-1}|| < \varepsilon$$

Линейная регрессия

$$Q(w) = \frac{1}{l}\sum_{i=1}^{l}(< w, x > \ -y_i)^2$$

$$\frac{\partial Q}{\partial w_1} = \frac{2}{l}\sum_{i=1}^{l} x_{i1}(< w, x > \ -y_i)$$

$$\ldots$$

$$\frac{\partial Q}{\partial w_d} = \frac{2}{l}\sum_{i=1}^{l} x_{id}(< w, x > \ -y_i)$$

$$\nabla Q(w) = \frac{2}{l}X^T(Xw \ - y)$$

Сложности градиентного спуска

- Для вычисления градиента, как правило, надо просуммировать что-то по всем объектам

- И это для одного маленького шага!

Оценка градиента

$$Q(w) = \frac{1}{l}\sum_{i=1}^{l} L(y_i, a(x_i))$$

- Градиент:

$$\nabla Q(w) = \frac{1}{l}\sum_{i=1}^{l} \nabla L(y_i, a(x_i))$$

- Может, оценить градиент одним слагаемым?

$$\nabla Q(w) \approx \frac{1}{l}\sum_{i=1}^{l} \nabla L(y_i, a(x_i))$$

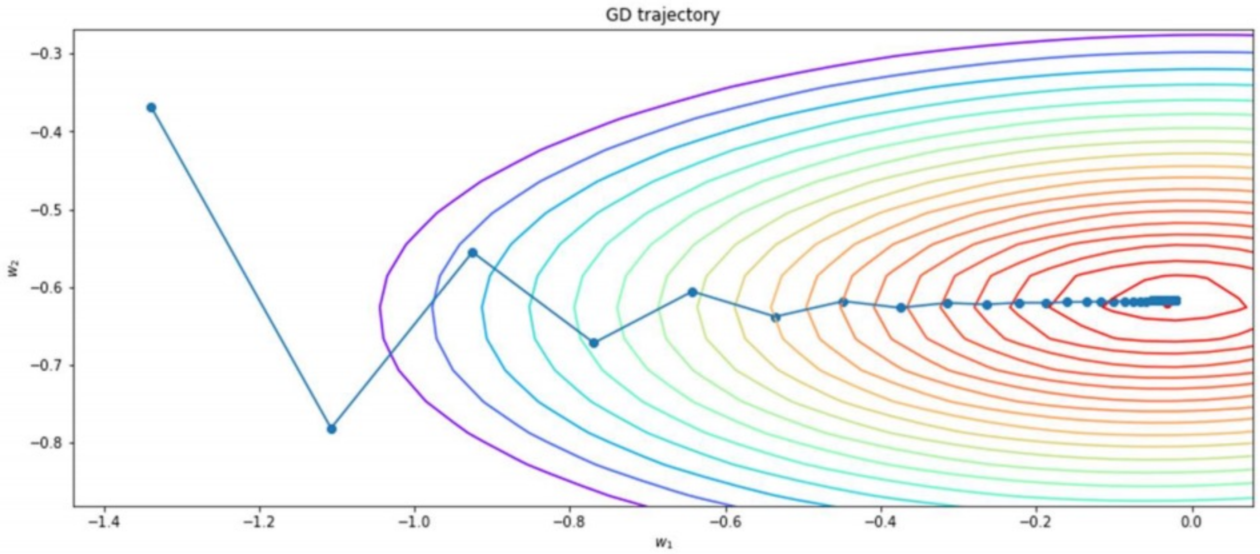
Стохастический градиентный спуск

1. Начальное приближение: w^0

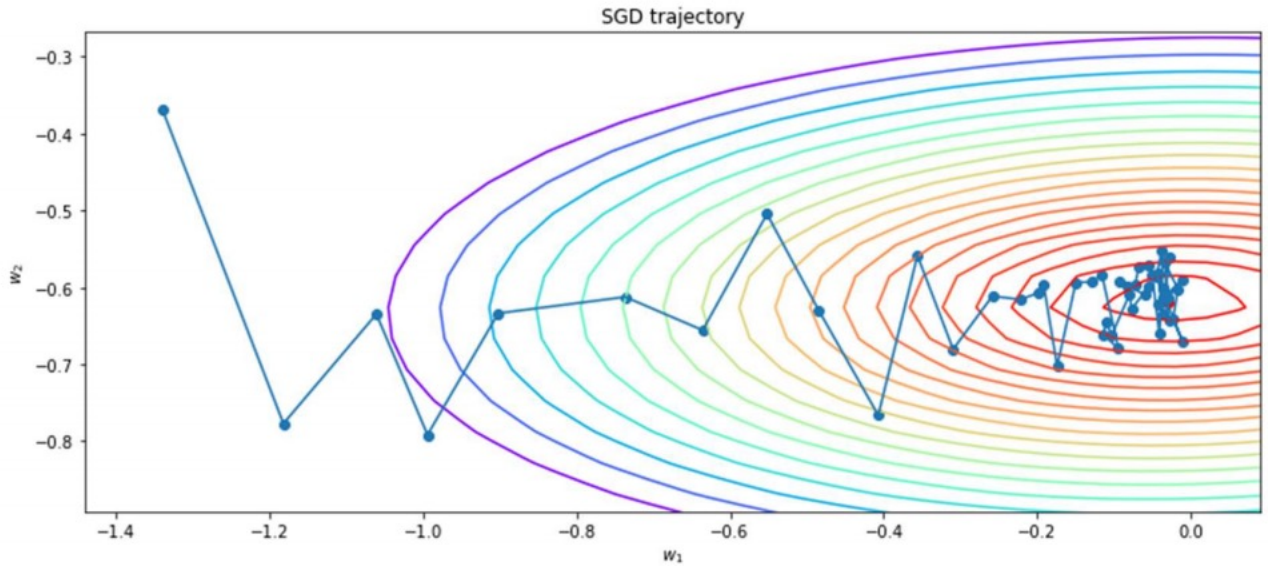
2. Повторять, каждый раз выбирая случайный обьект
   $i_t : w^t = w^{t-1} - \eta \nabla L\left(y_{i_t}, a\left(x_{i_t}\right)\right)$

3. Останавливаемся, если
   $$\left|\left|w^t - w^{t-1}\right|\right| < \varepsilon$$

Градиентный спуск



Стохастический градиентный спуск



Стохастический градиентный спуск

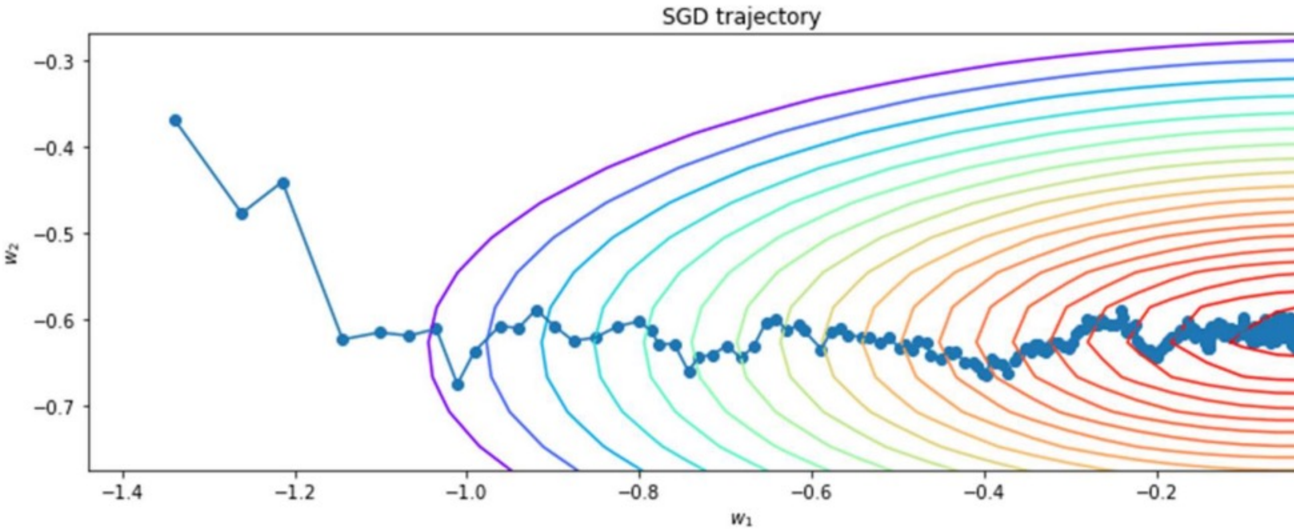
1. Начальное приближение: w^0

2. Повторять, каждый раз выбирая случайный обьект $i_t : w^t = w^{t-1} - \eta_t \nabla L(y_{i_t}, a(x_{i_t}))$

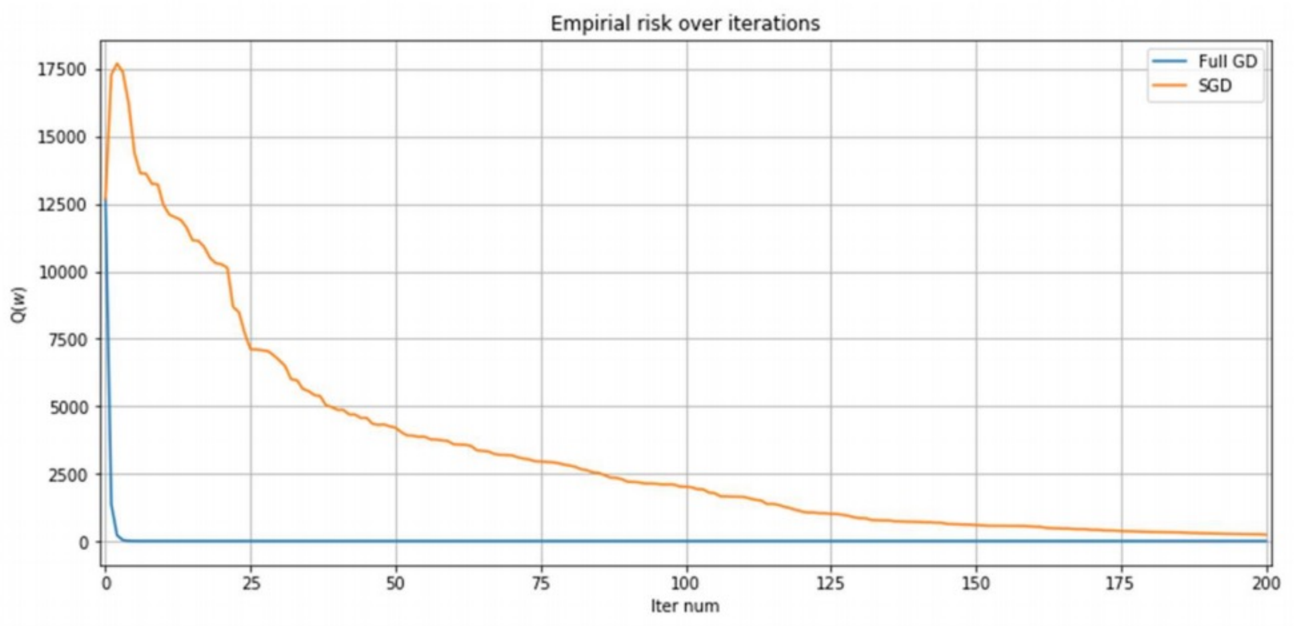
3. Останавливаемся, если
$$||w^t - w^{t-1}|| < \varepsilon$$

Стохастический градиентный спуск

$$\eta_t = \frac{0.1}{t^{0.3}}$$



Стохастический градиентный спуск

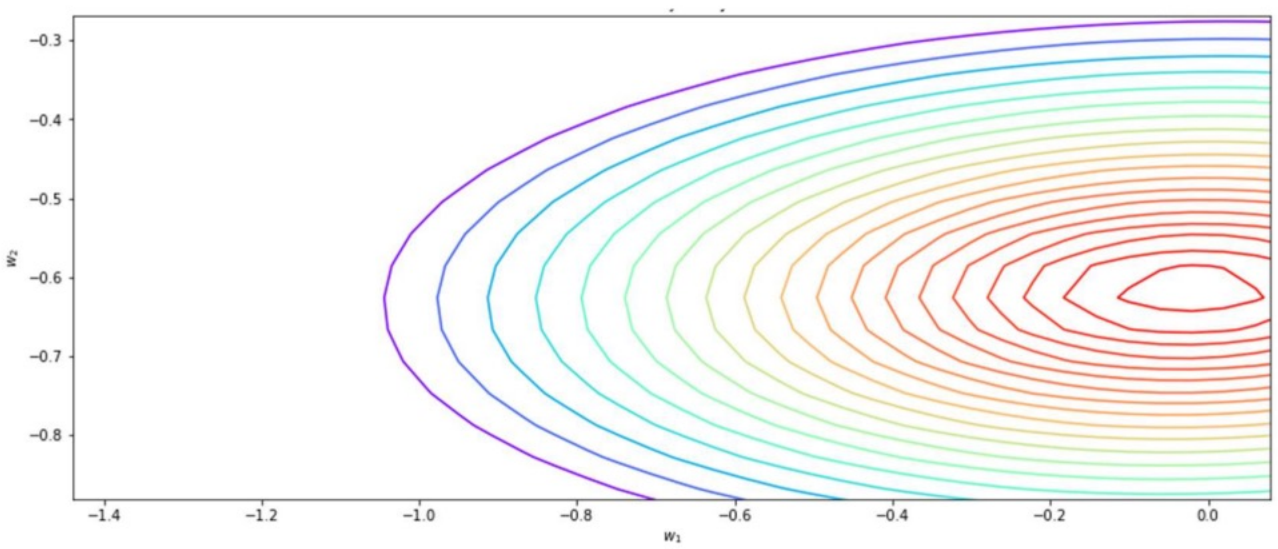


Стохастический градиентный спуск

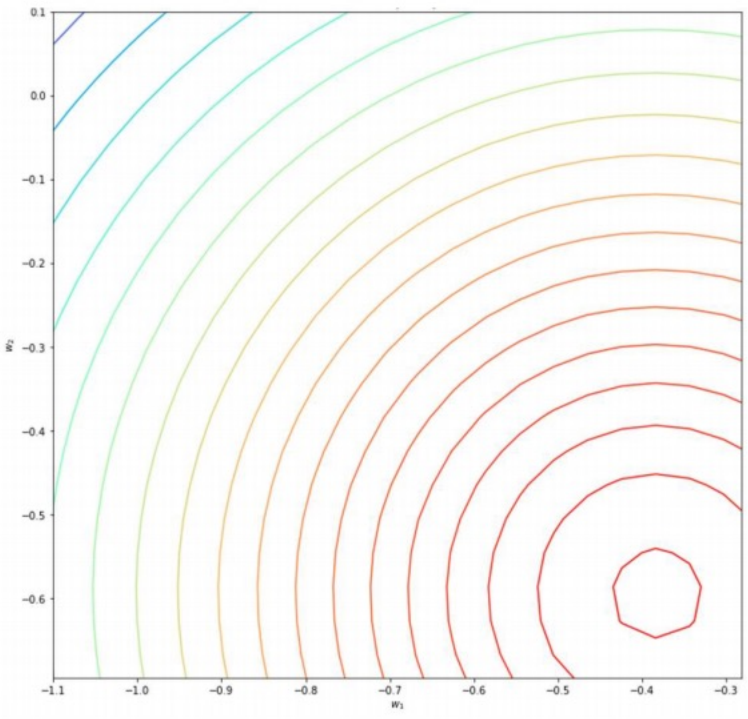
1. Начальное приближение: w^0

2. Повторять, каждый раз выбирая m случайный обьект $i_1, \ldots, i_m$:
   $w^t = w^{t-1} - \eta_t \frac{1}{m}\sum_{j=1}^{m} \nabla L(y_{i_j}, a\left(x_{i_j}\right))$
3. Останавливаемся, если
$$\left||w^t - w^{t-1}|\right| < \varepsilon$$

Масштабирование признаков



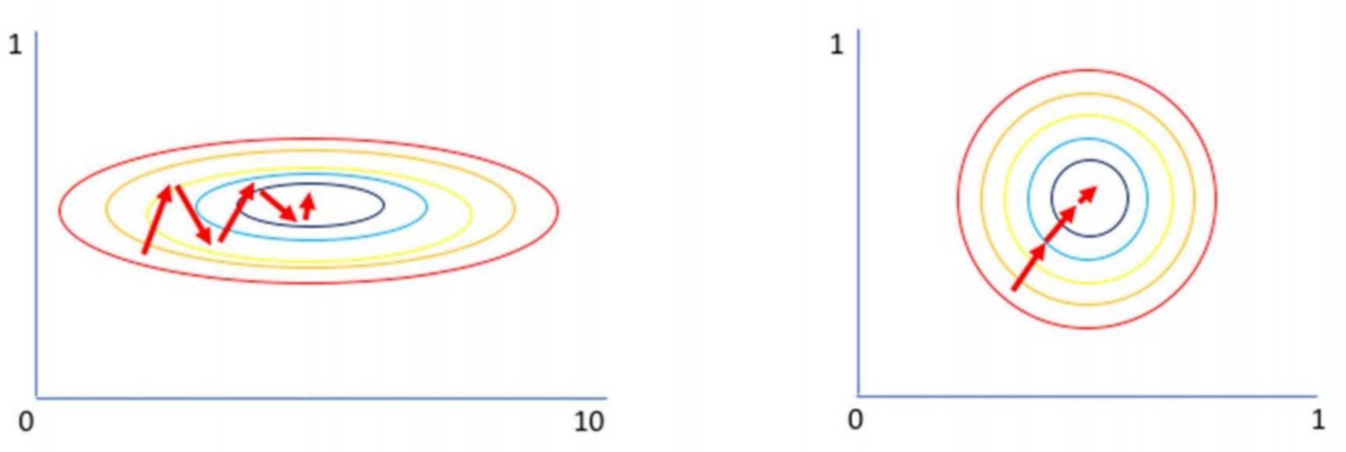
Масштабирование признаков



Масштабирование признаков

- До масштабирования значения градиента, соответствующие большим признакам, преобладают над остальными

- После масштабирования все параметры обновляются в равных пропорциях



Спасибо за внимание!

Ildar Safilo

@Ildar_Saf
irsafilo@gmail.com
https://www.linkedin.com/in/isafilo/